View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UUM Repository

Batch Sizing and Lot Splitting Strategies to Reduce Cycle Time in Semiconductor Assembly Process

Man, Siti Mariam[1, a], Zain, Zakiyah[1, 2, b], Nawawi, Mohd Kamal Mohd[1, c]

[1]School of Quantitative Sciences, College of Arts and Sciences,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.
[2]Centre for Testing Measurement and Appraisal (CeTMA),
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.

[b)]Corresponding author: zac@uum.edu.my
[c)]mdkamal@uum.edu.my

***Abstract*— Increasingly, products with customized features are manufactured in small batches based on demand by unique customers. Such diversity leads to high product mix, thus severely impacting both production planning and scheduling activities such as frequent equipment setup and conversion. Despite increased complexity, production continues being pressured on minimal cycle time to meet on-time delivery requirements. Focusing on the most challenging process, which is assembly, this paper presents batch sizing and lot splitting strategies to improve production cycle times. In the event of too short cycle time, overlapping technique is the most applicable, while "parallel" technique applies to normal cycle time. The commonly known "batch & queue" technique works well for long cycle time such as new product qualification lot.**

***Keywords*—** *cycle time, batch sizing, lot splitting and production planning.*

1. Introduction

Semiconductor environment consistently focuses on small batches production due to the need for the products to reach market in a shorter lead time. To fulfill this requirement, capacity planning model is designed with the flexibility of lot splitting to accommodate the speed of lot movement in production floor. The main advantage of lot splitting is that in practice, lots of small sizes can start its downstream operations earlier than larger lots. As smaller lot sizes continue gaining popularity over the past decades, lot splitting proves to accomplish the benefits of small batches even in industries with large customer orders. In other words, the small lots during production can later be merged for large quantity shipment; split and merge is a norm in production.

A vast literature on operations management has established that batch sizing decisions affect performance indicators such as flow times and work-in-progress (WIP) levels [1]. Gomes et al. [2] urged these batch sizing decisions are determinants of the responsiveness of any production environment. Thus, setting of batch sizes in a production system is indeed a critical control., as reported by Hopp et al. [3]. On the other hand, Jacobs et al. [4] acknowledged that splitting into small batch sizes will increase in number of equipment setups and conversion, thus directly impacting the optimum equipment efficiency (OEE).

When assessing the impact of batching decisions, it is important to distinguish between process batches and transfer batches [5]. A process batch size refers to the quantity of a product produced between two consecutive setups. In multi-product environment, process batching is often necessitated by changeover time which consumes part of the machine capacity [6]. The impact of process batching on the performance of production system performance was investigated in-depth by Karmarkar [7]. It was reported that long lead times impose costs attributed to higher WIP inventory, larger safety stocks and poorer performance to committed lead-time. Later, Nieuwenhuyse et al. [8] concluded that lot splitting is advantageous in reducing flow times and lowering congestion in production. Meanwhile, Chaharsooghi [9] urged that the kanban and delivery batch size are

International Journal of Supply Chain Management
IJSCM, ISSN: 2050-7399 (Online), 2051-3771 (Print)
Copyright © ExcelingTech Pub, UK (http://excelingtech.co.uk/)

critical to minimize total cost of quality if subject to scrapping. Chiu et al. [1] applied a mathematical model to derive the optimal manufacturing batch size for small batch production, while Chien [10] developed an adaptive model to rapidly respond to change in batch sizing to production line status. Gomes et al. [2] reported that batch sizing is a key decision to do by manufacturing in order to determine desired cycle time. The ideal batch sizing is that suits maximum of a machine capacity a day. Almost all the researchers concluded that smaller batch sizing gives a significant impact to shorten the process at each step, resulting in shorter lead time. The linearity of the material flow is key to gaining speed of execution. Any setup incurred would jeopardize the equipment OEE, resulting in slower material flow to next process.

Meanwhile, a transfer batch refers to the size of a sublot of the process batch, moved from one operation to another [5]. The transfer batching is driven by flow consideration; the smaller transfer batch allows subsequent operations to overlap, and thus reduces flow times by smoothing workflow and minimizing congestion levels [11]. Umble et al. [12] recommended synchronous flow as solution of choice in dealing with productivity problems, whereas Nieunhuyse et al. [6] examined the impact of transfer batch through an overlapping operation via sublot to gain productivity. Azizi et al. [13] discussed the workload leveling function for flawless execution in manufacturing, while Quadt et al. [14] outlined how to de-bottleneck the lot sizing and scheduling to smoothen assembly processes. In summary, transferring a smaller batch is always faster comparing to a bigger batch which need to have more than one equipment to be added in order to maximize the throughput within shorter cycle time. Figure 1 illustrates the transfer batching policy used in this study.

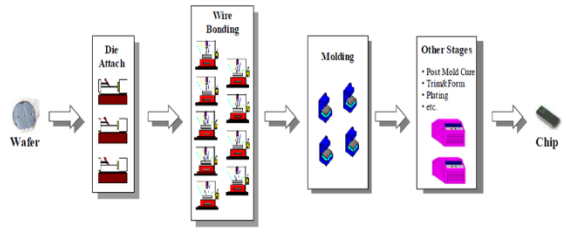

Figure 1. Transfer batching policy illustration (adopted from Quadt [14])

Figure 1 shows the processes involved in assembling the material from wafer level until chip. The wafer is diced and transferred onto lead frame at die-attach process, before moving to wire bonding process. Once completed, the material flows to molding process and then various processes: post mold cure, trim and form, plating, singulation and packing. The lots are split into smaller batches in sublot basis to reduce the processing hours at constraint operations and further split on needed basis to speed up the process in order to minimize cycle time. However, the batch will be merged back to the sublot at molding process and purely reverted to the original batch size prior to ship back to customers. The goal is to have a smooth flow for the batch to complete the entire processes in assembly production.

The lot splitting policy is also referred as overlapping operations principle [6], where lot quantity is equally split to run parallel on more than one equipment. The impact of lot splitting in deterministic production line environments has been studied by many researchers. Jacobs et al. [4] suggested a method to invoke an attempt with extra setups upon lot splitting, while Kalir et al. [15] provided lot splitting format for manufacturing guidance. Nieuwenhuyse et al. [6, 16, 8] discussed how the policy for transfer batch sizing affect the average transfer batch lead time and the total gap time in the production system. Meanwhile, Bukchin et al. [17, 18] investigated the lot splitting scheduling problem in a two-machine flow-shop environment with detached setups and batch availability. A manufacturing intelligence approach was developed by Chien et al. [10]. They formulated a basic model to predict the cycle time of the production line via integration of Gauss-Newton regression method and back-propagation neural network. Gomes et al. [2] advocated lot splitting strategies resulting in effective production and materials flow control whereby splitting lots to be processed on two equipments instead of one, leads to 50% cycle time reduction.

This paper begins with an overview in Section 1 and applicability of three lot splitting strategies to apply in semiconductor manufacturing as described in Sections 2 and 3. These sections provide some highly relevant insights into the effect of lot splitting on the behaviour of the subsequent stage to prepare for the arrival lot to that process. The effects on production run start times and process batch splitting flexibility, are further elaborated in Sections 4 and

5, while Section 6 summarizes the main results and conclusions.

2. Batch Sizing

Capacity-based lot-sizing is resource-specific, which is based on the total capacity of the available resources and how that capacity is consumed to run production line [19]. It assumes that in order to increase throughput, there are two main aspects to focus: reducing the inventory and minimizing the operating cost. In addition to reducing inventory, capacity-based lot-sizing increases throughput by splitting technique at the constraint operations. It uses only the existing capacity which does not increase the operating expense associated with machine and labor capacity. Capacity consists of a few components as illustrated in Table 1 below.

Table 1. Capacity components (Enhanced from Don Guild, Synchronous Management Consultant)

1) TOTAL DAILY CAPACITY		
2) UPTIME PERCENTAGE		5) DOWNTIME
3) CHANGEOVER TIME AVAILABLE	4) CYCLE TIME REQUIRED	

Referring to Table 1, the total capacity-based component (1) consists of equipment uptime (2), which minus out the downtime (5) and change over time available (3) in order to determine the cycle time required (4). While the first two components of changeover time available and machine down time are known and the third of cycle time required is a move time or in common word as throughput is somewhat easy to determine especially when material handling systems are not the bottlenecks. However, during running production within the cycle time required capacity, the queue time happens whenever a lot is staging on rack waiting for machine to be available. Besides, certain operation lot is waiting for people to inspect or manual move to next process. The resource availability is, in-turn, a function of the demand and the scheduling strategy [20].

Reducing the batch size can improve the manufacturing lead time. However, each processed piece requires the same amount of time to complete the processes, regardless of the batch size. The answer to lead time then lies in the queue time, also known as non-instant availability [12]. The manufacturing lead time can be broken down into various parts: set-up time, run time, move time, and queue time [21]. Queue time is usually much larger than the sum of the other numbers. The only number that considers the size of the order is the run time. The run time can be divided into process time and wait time for each individual piece in the batch as shown in Table 2 below.

Table 2. Manufacturing lead time components (adapted from [21])

Manufacturing Lead Time	
Mother Lot Size	Sublot Size
Setup Time	
Run Time	Process Time
	Wait Time
Move Time	
Queue Time	

In general, the wait time for each lot may still be long, despite a short queue time. It is usual for the first lot to wait after being processed; wait time consumes most of the run time, while the other lots are being processed. As for the last lot, most of the run time is spent waiting before being processed at the next operation. Meanwhile, those lots in the middle of the batch often spends equal amounts of time waiting before and after being processed. It is to be noted that there is a substantial amount of waiting in a queue during run time as estimated in an example shown in Table 3, and Table 4 provides its summary of run time and queue time.

Table 3. Queue Time Estimation

Mother Lot A (30K)	Run Time Process A (Hours)	Queue Process A (hours)	Run Time Process B (Hours)	Queue Process B (hours)	Equip #
Sublot A1 (5K)	0.75	0	20	0	Equip X
Sublot A2 (5K)	0.75	0.75	20	20	
Sublot A3 (5K)	0.75	1.5	20	0	Equip Y
Sublot A4 (5K)	0.75	2.25	20	20	
Sublot A5 (5K)	0.75	3	20	0	Equip Z
Sublot A6 (5K)	0.75	3.75	20	20	
	4.5	**11.25**	**120**	**60**	

Table 4. Runtime versus Queue time

Processing Hours	Hours Taken	Total
Total Run Time	4.5 + 120	124.5
Total Run Time with Splitting	4.5 + 20	24.5
Total Queue Time	11.25 + 60	71.25
Total Queue Time with Parallel Run	11.25 + 20	31.25

The queue time happens when the batch size is planned to be processed on one or minimum equipment. As shown above, process A run time is only 4.5 hours per sublot but has incurred 11.25 hours of queue time. Meanwhile, at process B, with the splitting into three equipment X, Y and Z, to parallel run, queue time is only 20 hours versus 120 hours if running on a single equipment. There is much gain to plan correctly on whether to run on big batch or splitting mode.

3. Transfer Batch Techniques

There are three mechanisms in transfer batching throughout the entire assembly line through batch and queue, parallel and overlapping techniques. The batch and queue technique involve sending the lot together as a group through each operational step [14], assuming each operation taking different timing based on the complexity of the process. Figure 2 shows seven operations to be completed sequentially in forty hours.

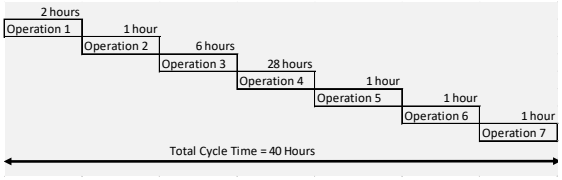

Figure 2. Batch and Queue Technique

Operation 4 shows the longest processing time followed by operation 3, while the rest of the operation is at the minimal processing time. Conventionally, industries with crucial fast response time have fixated on producing limited product variety, whereby inventories are built ahead of demand to enable immediate shipment upon receiving customer orders [22]. This technique is best for build to stock (BTS) environment. BTS does not require cycle time measurement. However, building to stock becomes costly and impractical when the number of products is high. In cases when demand is stochastic or negatively correlated among products, BTS imposes high risk to producer. In addition, BTS requires a huge warehouse to store finished goods which poses risk of being obsolete if market demand drops [1].

Parallel technique [13] uses the splitting mode of the batch to move with a smaller batch size at bottleneck operations in order to minimize processing time, as illustrated in Figure 3.

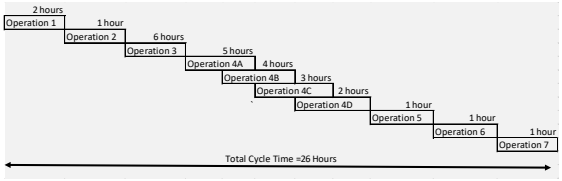

Figure 3. Parallel technique

This technique is more efficient in reducing total cycle time, by creating a smaller sublot size within a batch itself [24]. Assuming in this batch is consisting of total 4 lots. Upon completing the first lot of the batch at operation 3, the lot can move to the next process at operation 4. However, since operation 4 is the most constraint process, the lot can be split into sublot basis; a total of 4 equipments need to be used to produce the 4 sub-lots with 1 sub-lot on each equipment. Obviously, this technique requires multiple setups. Furthermore, this technique will only incur the longest setup at operation 4A as the first equipment setup. Upon setup completion at Operation 4A, recipe management system (RMS) is established and proliferated to other Operations 4B, 4C and 4D. RMS helps to reduce setup time for the subsequent equipment since the proliferation method used. However, the first equipment setup will take longer in order to develop the recipe. During sub lot basis, this method would be able to reduce the overall processing time through the right recipe proliferation application [24]. Parallel technique is the most popular technique in semiconductor manufacturing nowadays. It gives advantages to build to order (BTO) environment as urged by Yadav [23] from both aspects of shorter cycle time and cost measures. It requires a close monitoring with system flexibility and visibility on the lot splitting. With recent RMS in place, all operations are managed by RMS database.

Meanwhile, overlapping technique is the most efficient technique where equipment setup for the most constraint operation is accomplished ahead [25] (Simon, 2012). This strategy requires a pre-plan dummy process for offline setup as depicted in Figure 4.

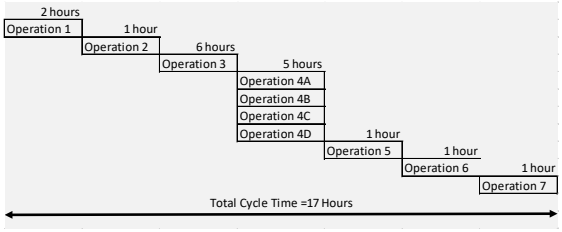

Figure 4. Overlapping technique

This overlapping technique as portrayed in Figure 4 requires an additional capacity where the equipments are prepared prior lot arrival – they are setup ahead by using dummy units. The focus is at the bottleneck process, at operation 4, 4 equipments readily setup and to run parallel upon physical lot arrival. In semiconductor industry, this technique is used to manage the hot lots that need a shorter cycle time where they are urgently needed for line down situation. However, it is not practical for mass production lot as it incurs additional capacity where the equipments will be idling waiting for lot which translates into cost and equipment efficiency level.

Table 5 compares among the process batches technique based on the cycle time, setups, productivity, headcount and specific advantages.

Table 5: Comparison among process batches technique.

Batch and Queue	**Parallel**	**Overlapping**
Longest cycle time based on the queue.	Medium due to splitting of lot into a smaller size.	Shortest since the setups are done ahead.
Lowest constraint as setup only been done once to run continually for the whole batch.	Medium constraint as elapse setups incurs for splitting purposes.	Setups done ahead. Highest cost, but the lot flows efficiently.
Highest productivity.	Medium productivity as a few machines to be setup to run the whole batch to faster processing at constraint process.	Lowest productivity measured since machines are setup ahead which cost capacity lost to the OEE.
Lowest cost measures from headcount basis.	Medium cost since splitting require a few headcounts to setup a few machines.	Highest cost for additional headcount to offline setup machines, while the existing headcount is focusing on the running production lots.
Advantages to build-to-stock (BTS) business environment since this business strategy does not concern about the assembly cycle time.	Advantage for build-to-stock (BTS) business environment as to shorten assembly cycle time to complete assemble of the product.	Advantage to quick-turn product such as the end for product for showcase, customer lines down or special request for fast turnaround.

In real practice by semiconductor companies, reducing process batch size is ideally not to incur high setups, while improving overall cycle time. Reducing transfer batch size indicated as splitting, is to quick start the sublot at the next process while waiting for total batch to complete the current process. Moreover, reducing process batch size is the key driver in efforts to minimize finished goods inventory, forecast dependency and ship back finish goods to customers. Nevertheless, reducing process batch size may increase set-up frequency leading to lower productivity on the constraint process. Additionally, the companies with the largest inventories of BTS items are most likely to benefit from reduced finished goods stock.

4. Results

The implementation of the right planning to the right method will improve cycle time at optimum capacity. Based on data analysis, below is the optimum ratio from six million daily loading as shown in table 6 below

Table 6: Summary of cycle time scores

Batch and Queue	Parallel	Overlapping	Cycle Time
86.006%	13.968%	0.026%	5.2
67.090%	32.784%	0.126%	5.8
56.622%	43.239%	13.900%	7.7
32.35%	62.832%	4.818%	5.7

The correlation among the three techniques are:

The higher overlapping technique need to balance up with higher parallel technique. However, the overlapping technique should not exceed 5% of the entire WIP planning.

5. Summary and conclusion

The use of process batching is induced by capacity considerations at the constraint process, driven by the necessity of performing equipment setups during product switching. Three batching techniques are opted depending on available total time allocated as well as equipment availability. The batch and queue technique is the most applicable to build-to-stock environment, while parallel technique is the most applicable to build-to-order with high mix and low volume environment. The overlapping technique best suits hot lot planning. The impact of batch sizing is vital where the smaller is better to gain better cycle time. Depending on due date set into the system, planning of the technique to be used to move the total batch needs to be done upfront. The planning system should be able to categorize and opt

the selected technique based on lead time setting, available equipments and manpower for additional setup incurs as well as remaining time to shipment date. This article consolidates all the three batch sizing and lot splitting strategies, and detail out for best case option setting to dispatch the right lot by using the right technique. Upon three years of implementation, 50% reduction in cycle time has been shown which is beyond the expectation. We have gain 22% reduction in the first year, 16% in second year and finally achieve 50% of overall from 12 days cycle time to 6 days cycle time. This proves the desired cycle time shall be achieved with the right planning and execution.

6. Acknowledgement

The authors express gratitude to Universiti Utara Malaysia for support throughout this study.

References

[1] Chiu, Y.S., Lin, H.D., Hwang, M.H., and Pan, N., (2011). Computational Optimization of Manufacturing Batch Size and Shipment for an Integrated EPQ Model with Scrap. *American Journal of Computational Mathematics*, doi:10.4236/ajcm.2011.13023, 202-207.

[2] Gomes, C., Ribeiro, A., Freitas, J., and Dias, L. (2016). Improving Production Logistics Through Materials Flow Control and Lot Splitting. *7th International Conference*, ICCL 2016, Lisbon, Portugal, September 7-9, 2016, Proceedings (pp.443-453). DOI: 10.1007/978-3-319-44896-1_29

[3] Hopp, W.J., Spearman, M.L., and Woodruff, D.L. (2000). Practical Strategies for Lead Time Reduction, *Manufacturing Review* 3, 78-84.

[4] Jacobs, F.R. and Bragg, D.J., (1988). Repetitive lots: flow-time reductions through sequencing and dynamic batch sizing, *Decision Sciences* 19, 281-294.

[5] Kropp, D., and Smunt, T. (1990). Optimal and Heuristic Models for Lot Splitting in a Flow Shop. Decision Sciences 21(4):691 – 709. Doi.org/10.1111/j.1540-5915.1990.tb01244.x.

[6] Nieuwenhuyse, I. and Vandaele, N. (2004). Determining the optimal number of sublots in a single-product, deterministic flow shop with overlapping operations. *International Journal of Production Economics;* 92, 221-239.

[7] Karmarkar, U. (1987). Lot sizes, lead times and in-process inventories. *Management Science* 33, 409-423.

[8] Nieuwenhuyse, I., Vandaele, N., Rajaram, K. and Karmarkar, U. (2007). Buffer sizing in multi-product multi-reactor batch processes: impact of allocation and campaign sizing policies. *European Journal of Operational Research* 179, 424-443.

[9] Chaharsooghi, S.K., and Sajedinejad, A. (2010). Determination of the Number of Kanbans and Batch Sizes in *JIT Supply Chain System.* Vol. 17, No. 2, pp. 143-149

[10] Chien, C., Hsu, C., and Hsiao, C. (2012). Manufacturing intelligence to forecast and reduce semiconductor cycle time. *Journal of Intelligent Manufacturing*, 23:2281-2294.

[11] Goldratt, E.M. and Cox, J. (1984), *The Goal*: A Process of Ongoing Improvement. Vol. 17, No. 2, pp. 143-149.

[12] Umble, M.M. and Srikanth, M.L. (1990) Synchronous Manufacturing Principles for World Class Excellence, South-Western Publishing Co., Cincinnati, OH.

[13] Azizi, A. and Persona, A. (2011). Lot splitting scheduling procedure for makespan reduction and machine capacity increase in a hybrid. The International Journal of Advanced Manufacturing Technology, Volume 59, Issue 5-8, pp 775-786. DOI: 10.1007/s00170-011-3525-x

[14] Quadt, D., and Kuhn, H. (2018). Production Planning in Semiconductor Assembly.

[15] Kalir, A.A. and Sarin, S.C. (2001). Optimal solutions for the single batch flow-shop lot streaming problem with equal sublots, *Decision Sciences* 32, 387-397.

[16] Nieuwenhuyse, I. and Vandaele, N. (2006). The Impact of Delivery Lot Splitting on Delivery Reliability in a Two-Stage Supply Chain, *International Journal of Production Economics* 104, 694-708.

[17] Bukchin, J., Tzur, M. and Jaffe, M. (2002). Lot splitting to minimize average flow time in a two-machine flow shop. *IIE Transactions* 34, 953-970.

[18] Bukchin, J. and Masin, M. (2004). Multi-objective lot splitting for a single product machine flowshop line. *IIE Transactions* 36, 191-202.

[19] Yusof, U.K, and Khalid, A. (2015). Dynamism in a Semiconductor Industrial Machine Allocation Problem using a Hybrid of the Bio-inspired and Musical-Harmony Approach. *International Conference of Material Sciences and Engineering*.

[20] Wu S., and Onari H. (2016). Production Capacity Planning with Market Share Expansion under Different Procurement Strategies. *International Journal of Economic Management* 5, 345. doi:10.4172/2162-6359.1000345.

[21] Schragenheim, E., Dettmer, H., and Patterson, W. (2000). *Supply Chain Management at Warp Speed: Integrating the System from End to End.* ISBN 9781420073355, CRC Press Taylor and Francis Group.

[22] Gupta, D., and Benjaafar, S. (2010), Make-to-order, Make-to-stock, or Delay Product Differentiation?

[23] Yadav, A., Kumar, A., Raut, N., and Yadav, M. (2016). Optimization of Batch Volume in a Multi-Part Manufacturing System. *International Journal of Advanced Research in Science, Vol. 3, Issue 3, March 2016.*

[24] Saluja, V., and Jain, A. (2014). Optimization of Flexible Flow Shop Scheduling with Sequence Depent Setup Timw and Lot Spliting. *5th International & 26th All India Manufacturing Technology, Design and Research Conference* (AIMTDR 2014) December 12th–14th, 2014, IIT Guwahati, Assam, India.

[25] Simon, J., Mark, E., Mwangola, W., and Burke, G. (2012). Conditional Lot Splitting to Avoid Setups While Reducing Flow Time. *American Journal of Operations Research,* 2012, 2, 453-466. http://dx.doi.org/10.4236/ajor.2012.24054